A Method for Localizing Non-Reference Sequences to the Human Genome

Brianna Sierra Chrisman[1*], Kelley M Paskov[2], Chloe He[2], Jae-Yoon Jung[3], Nate Stockham[4], Peter Yigitcan Washington[2], Dennis Paul Wall [2, 3 *]

Departments of Bioengineering[1], Biomedical Data Science[2], Pediatrics[3], and Neuroscience[4],
Stanford University
Stanford, CA 94305, USA
Email: briannac@stanford.edu,dpwall@stanford.edu

As the last decade of human genomics research begins to bear the fruit of advancements in precision medicine, it is important to ensure that genomics' improvements in human health are distributed globally and equitably. An important step to ensuring health equity is to improve the human reference genome to capture global diversity by including a wide variety of alternative haplotypes, sequences that are not currently captured on the reference genome. We present a method that localizes 100 basepair (bp) long sequences extracted from short-read sequencing that can ultimately be used to identify what regions of the human genome non-reference sequences belong to. We extract reads that don't align to the reference genome, and compute the population's distribution of 100-mers found within the unmapped reads. We use genetic data from families to identify shared genetic material between siblings and match the distribution of unmapped k-mers to these inheritance patterns to determine the the most likely genomic region of a k-mer. We perform this localization with two highly interpretable methods of artificial intelligence: a computationally tractable Hidden Markov Model coupled to a Maximum Likelihood Estimator. Using a set of alternative haplotypes with known locations on the genome, we show that our algorithm is able to localize 96% of k-mers with over 90% accuracy and less than 1Mb median resolution. As the collection of sequenced human genomes grows larger and more diverse, we hope that this method can be used to improve the human reference genome, a critical step in addressing precision medicine's diversity crisis.

Keywords: alternative haplotypes, genomics, genetic diversity

Background

The Beginning: The Human Genome Project

In 2001, scientists announced the completion of the human genome sequence [1, 2]. This sequence eventually evolved into what is known as the "Human Reference Genome," the most recent stable version being HG38, and is used extensively in human genetics research and precision medicine. However, despite the many improvements to the human reference genome over the past two decades [3, 4], several issues to the human reference genome remain. First of all, contrary to the public's understanding of the Human Genome Project, the sequence of even a single human reference genome remains incomplete. Many hard-to-sequence or hard-to-

[*] Corresponding authors

assemble regions such as repetitive regions, heterochromatin, and much of the Y chromosome are still missing from the human reference genome [5]. Newly available long-read sequencing technologies are promising avenues for mitigating these issues, and the Telomere-to-Telomere Consortia recently published the first fully sequenced human chromosome (chromosome X) [6], with the rest of the chromosomes soon to follow.

The other pressing issue, and the one that we will focus on in this paper, is that, in its most commonly used form, the human reference genome is a linear sequence of bases built using the genomes from only a handful of individuals. While a linear reference genome is an acceptable reference for analyzing DNA that is fairly similar to the reference genome, it is extremely difficult to localize segments of DNA containing structural variations to a linear reference genome [7]. In many sequencing pipelines, reads that do not align well enough to the human reference genome get thrown out and go unused in downstream analysis [8]. A single linear reference genome has become so problematic that the human reference genome effort has already attempted to include several *alternative haplotypes*, sections of genome that are found in a nontrivial fraction of the population but that differ greatly from the reference genome [9].

The Current Crisis: Grappling with Missing Genetic Diversity

Unfortunately, the last decade of genetics has primarily sequenced individuals of European ancestry. Consequently, we know much more about the European-specific SNPs, structural variants, and alternative haplotypes than we do about genetic variations in any other ancestry [10]. Notably, African genomes have been particularly underrepresented in large-scale genomic studies, despite the fact that people with African ancestry have one of the most diverse collection of genomes in the world [11]

These alternative haplotypes may in fact play an important role in disease and precision medicine: a structural variant in KCNJ18, a gene with alternative haplotypes missing from most maps of the human genome, can cause a Mendelian neurological disease called thyrotoxic periodic paralysis, most common in Asian and Latin American men [12]. Models for dosing Warfarin, a blood thinner that is metabolized at a different rates based on patient genetics, have been built using primarily European genomic data on the VKORC1 and CYP2C9 haplotype groups, leading to inequities in Warfarin efficacy and safety [13]. With the field of genomics slated to make great leaps in precision medicine in the next decade, genomic researchers must address this diversity crisis as soon as possible [14].

The Future: Towards a More Diverse Human Reference Genome

One of the first steps in making the field of human genomics fair, equitable, and accessible to all is to improve the human reference genome in order to better represent the diverse set of DNA that represents the collective human genome. Already many large scale sequencing efforts have been launched, aimed at cataloguing genomes from diverse ancestries, particularly those who have been historically underrepresented in genomic studies [15–19]. Meanwhile, the field of pan-genomics has been searching for a way to structure and display the collection of all possible genomic sequences from humans [20], with graph genomes emerging as the most popular solution. A non-linear reference genome has become so popular that the Genome Reference

Consortium delayed their upcoming release of HG39, in order to consider alternatives to the traditional linear representation [21].

In order to go from a collection of many genomes, to a graph or other non-linear representation of the human pan-genome, the first step is determining where exactly alternative haplotypes, sequences not captured in the primary human reference genome, are coming from. That is, we need to be able to *localize*, or identify the chromosome and approximate location of, alternative haplotypes and non-reference sequences. Nearly all all localization of human alternative haplotypes has been done using long-read sequencing [22–24], which relies on sequencing stretches of tens of thousands of basepairs, with the hope of a read encompassing an alternative haplotype flanked by sequences of DNA that are easy to align to the reference genome, effectively determining where on the genome this uncatalogued sequence lies. However, long-read sequencing is still not nearly as popular as short read sequencing and is more error prone [25]. Most importantly, many of the recently launched studies aimed at underrepresented ancestries use short-read sequencing [15, 16, 19, 26].

We present a method that localizes 100-bp long k-mers extracted from short-read sequences. We use genetic data from over 700 families to identify shared genetic material among siblings and match the distribution of unmapped k-mers to these inheritance patterns to determine the most likely genomic region of a k-mer. We perform this localization with a computationally tractable Hidden Markov Model coupled to a Maximum Likelihood Estimator. We show that our algorithm is able to localize 100-mers with over 90% accuracy, and is capable of resolution <1Mb. As the collection of sequenced human genomes grows larger and more diverse, we expect this method will be of use to improve the human reference genome, the first step to addressing genomics' and precision medicine's diversity crisis.

Methods

Dataset

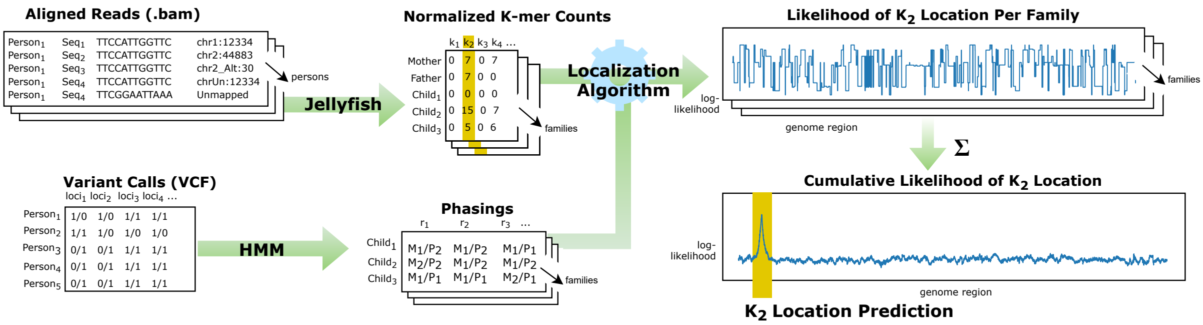
We used the iHART WGS collection [27], a dataset of multiplex autism families, containing 1,006 families and 4,610 individuals. Individuals were sequenced at 30x coverage using Illumina's TruSeq Nano library kits, reads were aligned to build GRCh38 of the reference genome and decoy contigs using bwa-mem [28], and variants were called using GATKv3.4. Only biallelic variants that passed GATK's Variant Quality Score Recalibration (VQSR) were included in analysis.

K-mer Counts

Our method uses distributions of k-mer counts within and across families. Thus, in order to localize each k-mer the first step of our algorithm is to extract the counts of k-mers for all samples in our dataset, as shown in the first step in Figure 1.

We use alternative haplotypes with known locations as a ground truth dataset to evaluate the performance of our method. The more recent versions of the human reference genome include a decoy reference genome, which includes several hundred contigs of alternative haplotypes (chr#_ID_Alt) annotated with their location with respect to the primary reference genome.

For each sample in our dataset, we extracted the reads that mapped to any of these alter-



$Person_1$	Seq_1	TTCCATTGGTTC	chr1:12334
$Person_1$	Seq_2	TTCCATTGGTTC	chr2:44883
$Person_1$	Seq_3	TTCCATTGGTTC	chr2_Alt:30
$Person_1$	Seq_4	TTCCATTGGTTC	chrUn:12334
$Person_1$	Seq_4	TTCGGAATTAAA	Unmapped

	k_1	k_2	k_3	k_4 …
Mother	0	7	0	7
Father	0	7	0	0
$Child_1$	0	0	0	0
$Child_2$	0	15	0	7
$Child_3$	0	5	0	6

	$loci_1$	$loci_2$	$loci_3$	$loci_4$ …
$Person_1$	1/0	1/0	1/1	1/1
$Person_2$	1/1	1/0	1/0	1/0
$Person_3$	0/1	0/1	1/1	1/1
$Person_4$	0/1	0/1	1/1	1/1
$Person_5$	0/1	0/1	1/1	1/1

	r_1	r_2	r_3 …
$Child_1$	M_1/P_2	M_1/P_2	M_1/P_1
$Child_2$	M_2/P_2	M_1/P_2	M_1/P_2
$Child_3$	M_1/P_1	M_1/P_1	M_2/P_1

Fig. 1. The general pipeline for localizing sequences, a likelihood-based algorithm that uses childrens; IBD computed from a HMM and k-mers distributions derived from raw sequencing reads. Note that the graphs shown in the figure are real results, generated using our algorithm on our data.

native haplotypes. We used the fast multi-threaded k-mer counter "jellyfish" [29] to extract and count the number of 100-mers within the decoy alternative haplotype contigs. We chose a k-mer length of 100, because with 150 bp reads at 30x coverage (1) there is a high probability that the full sequence of given k-mer will show up in at least one sequencing read of an individual if their genome does in fact contain it; (2) there is almost zero chance that every count of a k-mer contains an error in an individual; (3) 100 bp is long enough to correspond to a unique k-mer within the genome, except in highly repetitive regions; and (4) 100 bp is long enough to make *de novo* assembly of k-mers possible for future steps in alternative haplotype construction. In order to avoid k-mers due to sequencing errors, we ignored individuals with only a single instance of a given k-mer, and only included k-mers that were found in at least 2 samples.

Phasing Families

Phasing refers to the use of an individual's genetic data to determine which sequences or variants in their genome were inherited from their mother and which from their father. Families share genetic material. Each parent has two copies of each chromosome. During meiosis, these two copies are combined in large blocks to form a new chromosomal copy which is then inherited by the child. Our goal is to identify which parental copy was inherited by each child at every position of the genome. Using a hidden Markov model (HMM) , our phasing algorithm ultimately outputs for each region of the genome, whether a child inherited copy 1 or copy 2 of their mother's genome in the form of $(m_{1|2},\, p_{1|2})$

We phase families using the final variant calls from the iHART dataset. We use a hidden Markov model to describe a state space of inheritance. An HMM is defined by a state space, transition probabilities, and emission probabilities. In our case, the state space is the set of ways that the children can inherit the two maternal copies of the chromosome (m_1 and m_2) and the two paternal copies (p_1 and p_2). For example, for a family with two children (m_1p_1, m_2p_1) represents a state where the first child inherits parental copies m_1 and p_1 and the second child inherits parental copies m_2 and p_1. Since we are working with whole-genome sequencing data, we also include a hard-to-sequence region flag in our state space in order to detect and flag regions with many sequencing errors. The transmission probabilities in our model represent recombination events where the chromosome inherited by a child switches from one parental copy to the other. We use estimates of $1.39e^{-8}$ and $9.23e^{-9}$ for the probability of maternal and paternal recombination per base-pair [30]. The emission probabilities in our model represent the probability of sequencing errors. We estimate these probabilities directly from the genotype data using a family-based method [31]. Finally, we use the Viterbi algorithm [32] to identify the sequence of states that best explains the observed variant calls. The result is a fully phased family as shown in Figure 2.

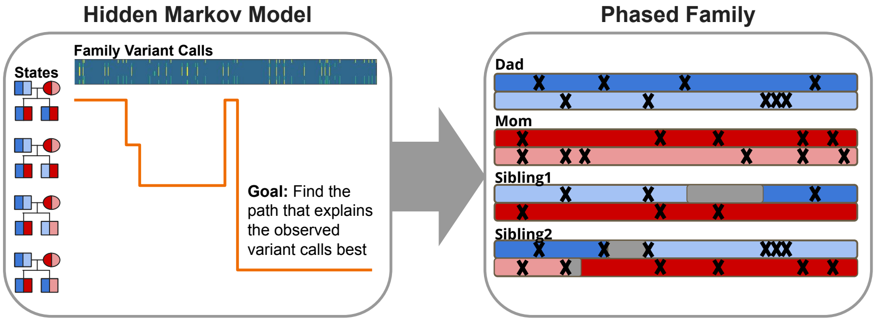


Fig. 2. Our hidden Markov model uses the pattern of variants within a family, to detect which parental copy was inherited by each sibling. This allows us to detect recombination points as well as identity-by-descent (IBD) sharing between the siblings.

Our phasing algorithm identified 225,313 total recombination points, and thus the same number of possible distinct regions to use in the subsequent maximum likelihood localization algorithm.

Note that while phasing families is an algorithmic challenge in its own right, we make our localization algorithm the main focus of paper.

Localizing Haplotypes

For each k-mer, we wish to find it's corresponding location (region r) in the genome that best explain the distribution of the k-mer counts in all of the families.

We define the distribution of a given k-mer in all samples as $\mathbf{K}$, and the distribution of a given k-mer in family f as $\mathbf{K_f}$. Therefore, we want to find the region r that maximizes the likelihood of observing $\mathbf{k}$. We can rewrite this likelihood in log-likelihood form:

$$\ell(r; \mathbf{K}) = \sum_f log(P(\mathbf{K_f}|r)) \tag{1}$$

Let's discuss how to compute $P(\mathbf{K_f}|r)$, the probability of a family's k-mer counts given the k-mer's hypothetical region.

We assume that children inherit the given k-mer in a Mendelian fashion: from each parent, they receive either 0 or 1 copy of the k-mer. A parent with the k-mer present on both copies of their chromosome is guaranteed to pass down the k-mer to their child, a parent without the k-mer present on either copy of their chromosome will never pass the k-mer down to their offspring, and a parent who is heterozygous for the k-mer has a 50% chance of passing the

k-mer down to an offspring. The possible phased genotypes denoted in the order of (maternal allele, paternal allele) for any person are: 0/0 if the person does not have the k-mer on either copy of their chromosomes; 1/0 if the person has the k-mer on the maternal copy (or copy 1) of their chromosome,but no the paternal copy (copy 2); 0/1 if the person has the k-mer on the paternal copy of their chromosome (copy 2), but not the maternal copy (copy 1); and 1/1 if the person has the k-mers on both copies of their chromosome.

We will refer to define this set of possible genotypes as $\mathbf{G} = \{0/0, 1/0, 0/1, 1/1\}$. We will call mother's genotype g_m, father's genotype g_p (p for paternal), and a child's genotype g_c. We can also define mother's, father's, and child's k-mer counts as k_m, k_p, and k_c respectively. Using this notation and the law of total probability, we can rewrite our family-wise probability of observing the data as:

$$P(\mathbf{K_f}|r) = \sum_{g_m, g_p \epsilon G} P(\mathbf{K_f}|r, g_m, g_p) P(g_m, g_p) \tag{2}$$

$$= \sum_{g_m, g_p \epsilon G} \left(P(k_p|g_p) P(k_m|g_m) \prod_c P(k_c|g_c) \right) P(g_m, g_p) \tag{3}$$

We iterate over all possible genotypes for the mother and the father but not for the children because from our phasing algorithm, we already know which copy of mom's DNA a child inherited, and which copy of dad's DNA a child inherited at any given region. We can therefore compute the child's genotype for a k-mer, given the region, and the parent's phased genotypes:

$$g_c = phase(c, r, g_m, g_p) \tag{4}$$

The $phase()$ function queries the inheritance pattern of a child c at a region r in the phasing dictionary. Using the phasing information, it then combines the maternal haploid genotype on the appropriate copy of g_m with the paternal haploid genotype on the appropriate copy of g_p to infer the child's diploid genotype.

Let's compute a toy example: $phase(c^*, r^*, 1/0, 0/1)$. If a phasing dictionary tells us that at region r^*, child c^* inherited DNA from her mother's chromosomal copy 1 and her father's chromosomal copy 2, the child would then inherit one copy of the k-mer from her mother (because the mother's genotype is 1/0, corresponding to a copy on the mother's 1 copy), and another copy of the k-mer from her father (because the father's genotype is 0/1, corresponding to a copy on the father's 2 copy). The child's genotype g_{c^*} would thus be 1/1. By the same logic, for the same child at the same region, $phase(c^*, r^*, 0/0, 1/0)$ would be 0/0.

Now let's compute $P(k, g)$, the probability of a k-mer count in a person, given a person's genotype. We assume that k-mers do not exhibit copy number variation; a given k-mer only appears once in each haploid copy of a person's genome. We also assume that the sequencing pipeline, which used random-PCR targeted every region of the genome at an equal read depth (we later discuss the limitations of these assumptions in the discussion). k-mer depth then follows a Poisson distribution, dependent on genotype heterozygosity and average k-mer depth μ_k. Using the syntax where $\sum(0/0) = 0$, $\sum(0/1) = \sum(0/1) = 1$, and $\sum(1/1) = 2$, a k-mer distribution can be summarized as follows:

$$P(k|g, \mu_k) = P_{\text{poisson}}(k; \mu_k \sum g) \tag{5}$$

A theoretical μ_k can be derived for each person using the total number of sequencing reads, the length of the person's genome (which differ slightly between males and females), and the length of the k-mer. The average μ_k of a 100-bp k-mer for our samples was 5.83.

Given the phasings and k-mer counts, for every family we can now compute the log-likelihood of a given k-mer belonging to each region of the genome $log(P(\mathbf{K_f}|r))$ and we can take the cumulative log-likelihoods to compute the total log-likelihood of a given k-mer belonging to each region of the genome $log(P(\mathbf{K}|r))$. Rather than reporting only the region with maximum likelihood and be at the whim of statistical noise, we estimate a maximum likelihood interval [33]. From our cumulative likelihoods, we find all the neighboring regions on the graph whose relative likelihoods are within a certain threshold. That is, our maximum likelihood interval is:

$$\{r : \frac{L(r)}{L(\hat{r})} >= t\} \tag{6}$$

where t is a certain threshold. Because k-mers very in their allele frequencies and families in which they are present, and because families vary in their IBD patterns, k-mer likelihood profiles are susceptible to different amounts of statistical noise. For that reason, we choose t as a function of the standard deviation (σ) in each k-mer's log-likelihood profile. Specifically, we define our maximum likelihood region to be:

$$\{r : log\frac{L(r)}{L(\hat{r})}) >= -\gamma\sigma(log(L(r)))\} \tag{7}$$

where γ is a hyperparameter that we must tune. We tried values of γ between .01 and 1, ultimately choosing .1 for a balance of sensitivity and specificity, localizing 96% of medium-prevalence (prevalence of .2-.8) k-mers, with a 90% accuracy and median resolution of 870Kb. If regions from multiple different chromosomes fell into our maximum likelihood region for a given k-mer, we considered that k-mer unlocalized. A schematic of this pipeline is shown in Fig. 1, with actual family-wise and cumulative log-likelihood graphs computed by our algorithm on a k-mer from the data.

The code for our localization algorithm can be accessed at `https://github.com/briannachrisman/alt_haplotypes`.

Results

Our algorithm ultimately localized 96% of medium-prevalence k-mers, with a 90% accuracy and median resolution of 870Kb.

Likelihood Threshold Tuning

In order to test the trade off between accuracy and resolution, we experimented with several different relative likelihood threshold cutoffs (as described in Eq. 7 to determine localized region. We found that using a threshold of $\gamma = .1$ provided a satisfactory balance of accuracy and resolution. More lenient cutoff values, such as $\gamma = .5$ resulted in higher accuracy, but a lower resolution, and more k-mers with multiple chromosomes falling into their maximum likelihood regions and thus considered unlocalized.

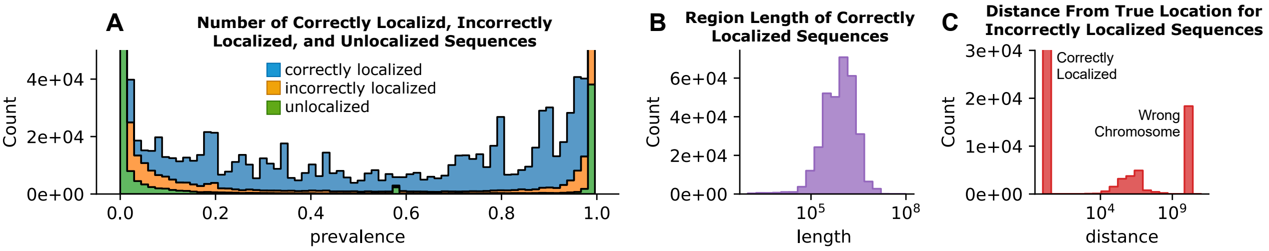


Fig. 3. (a) Number of sequences in correctly localized (colored as blue); incorrectly localized (yellow); or unlocalized (green). (b) Region length of correctly localized sequences. (c) Distance from true location for incorrectly localized sequences.

Dependence on Prevalence

As shown in Fig. 3A, localization capability and accuracy was heavily dependent on k-mer prevalence. Our algorithm was unable to localized or incorrectly localized k-mers with very high prevalence or very low prevalence. This is expected, as there will not be enough sibling discordance for a k-mer to determine which similarly inherited region between siblings it belongs to. Therefore, we report the % localized accuracy, median region length, and median distance from true values only for the k-mers with "medium" prevalences (defined as prevalences between between .2 and .8), and used those metrics to decide on the best likelihood cutoff.

Accuracy and Precision

Using a relative likelihood cutoff of -.1$\sigma(L)$, our algorithm was able to localize 96% of medium-prevalence k-mers from the decoy sequences, with 90% accuracy and a median 870Kb resolution. With a 3Gb long genome, this resolution corresponds to being able to localize a sequence to the a region .03% of the full genome length.

Error Profile

In order to understand the types of errors from our localization algorithm, we analyzed the error profile. From Fig. 4, we see that our errors are primarily generated from a handful of alternative haplotypes whose k-mers are incorrectly localized consistently. Only 24 out of 198 alternative haplotypes had their k-mers localized with < 90% accuracy. These alternative haplotypes include chr8_KI270813v1_alt, chr12_GL877875v1_alt, chr2_KI270894v1_alt , chr17_KI270907v1_alt, and chr22_KI270878v1_alt. Interestingly, for k-mers from many of these these 'poorly' localized haplotypes, our algorithm consistently localized the k-mers to the same region of the genome. For example, k-mers from chr12_GL877875v1_alt consistently were localized to the beginning of chromosome 6, and k-mers from chr8_KI270813v1_alt were consistently localized to a region 2Mb downstream of their annotated location. Given these patterns, we wonder if perhaps there are alternative haplotypes located elsewhere in the genome that are homologous to those in the decoy sequence and our algorithm is detecting such.

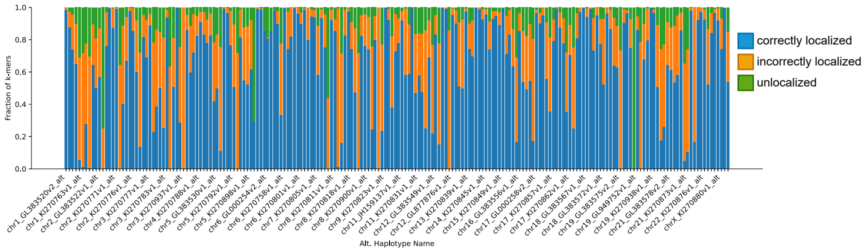


Fig. 4. Stacked histogram of correctly localized (colored as blue), incorrectly localized (yellow), and unlocalized (green) k-mers for each alternative sequence in the decoy genome.

As shown in Fig. 5, for the most part incorrect localizations are fairly randomly distributed across the genome. There are some hotspots for incorrect localization, such as the spikes in chromosome 6 and 8, which correspond to k-mers from a handful of alternative haplotypes being consistently localized to those regions. This mostly uniformly distributed error profile bodes well for the future next steps in constructing alternative haplotypes, which would be to *de novo* assemble k-mers that are localized to similar regions of the genome into longer contigs. A uniform error profile would likely mean that k-mers localized to the wrong region would simply not make it into the assembly of a long contig, and with a high enough k-mer overlap and accuracy a small percentage of k-mers missing from a region should not create gaps in the final assembly. We discus the potential pipeline downstream of our localization algorithm in more detail later on.

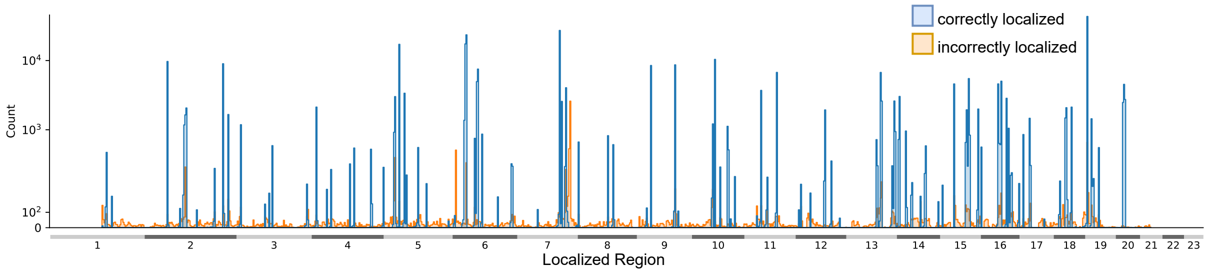


Fig. 5. Localization predictions for incorrectly and correctly localized k-mers.

Discussion

We show that our model can localize 96% of medium-prevalence k-mers extracted from alternative haplotype sequences, with over 90% accuracy and better than 1Mb resolution.

Use Cases

Our model can be used as the first step in a process to generate longer strings of sequences representing alternative haplotypes that can ultimately be used as alternate contigs in a linear reference genome, or nodes in a graph-based reference genome. In large genomic studies of nuclear families, we would extract k-mers from reads that didn't align to the reference genome, compute their counts for each person, then use our algorithm to localize each k-mer. From there, in a process similar to genome binning [34], k-mers that were localized to similar regions of the genomes would be *de novo* assembled into larger contigs, which could ultimately be used in future human reference genomes, linear or otherwise. Very near future work will involve localizing reads from our that did not map to the reference genome nor the decoy genome from our large dataset of families.

As mentioned before, most anticipated large-scale genomic studies focused on underrepresented groups will use short-read sequencing technology. We showed that our method works well for short-read sequences, but theoretically our method could be adapted to long-read sequencing as well. Given our read length and coverage, we chose a k-mer length of 100. In studies with different read lengths, coverage, and error rates, a different k-mer length may be optimal.

Limitations

This model has two major limitations. Firstly, in using a Poisson distribution to model k-mer count, we assumed that every 100-bp sequence occurs only once in the genome. This limits our ability to k-mers that might appear many times in a genome, such as those from highly repetitive sequences. However, ultimately we wish to assemble our localized sequences into longer contigs. Highly repetitive sequences are difficult to assemble, and may vary in number of repeats across individual. Our algorithm would likely still be able to localize the sequences corresponding to the ends of a highly repetitive contig and capture at least one full repeat, allowing for reads from a repetitive region to still map to such a contig.

Most importantly, our localization method relies on having genomes from many nuclear families. Though the field of genomics historically prioritized attaining data from families in order to understand inheritance patterns via pedigrees, recently family structure has been sidelined in favor of large unrelated case and control cohorts. However, genomic methods using families are extremely powerful and seem to be making a comeback. Using genomic data from families, we can not only make use of our described localization method, but we can better understand dynamics of recombination [35], detect rare and *de novo* variants and analyze their relationship to phenotypes [27], use sibling recurrence rates to categorize inheritance mechanisms of disease [36], and measure experiment-specific sequencing error rates [31]. We strongly advocate for the return of family study designs in the field of genomics, particularly in upcoming studies that seek to sequence large numbers of genomes from underrepresented groups.

The Role of Sequence Localization in the Future of Reference Genomes

Localization has an important role to play in the future of reference genomes. Being able to localize short sequences will allow us to construct alternative haplotypes from the unmapped read space of short-read whole genome sequences. Our algorithm uses nuclear family structure to localize such sequences, and we implore future studies to take advantage of our algorithm along with the the many other benefits that come with using family study design. Particularly, as the African Society for Human Genetics pushes forward its exciting goal of sequencing 3 million African genomes [37], we highly recommend recruiting families for at least some samples. A better understanding of diverse haplotypes, starting from where they are located within the human genome, is a vital first step in creating a reference genome that encompasses the full spectrum of human genetic diversity and in addressing precision medicine's diversity crisis.

References

[1] J. Craig Venter et al. "The sequence of the human genome". In: *Science* 291.5507 (2001), pp. 1304–1351. ISSN: 00368075. DOI: `10.1126/science.1058040`.

[2] Eric S. Lander et al. "Initial sequencing and analysis of the human genome". In: *Nature* 409.6822 (2001), pp. 860–921. ISSN: 00280836. DOI: `10.1038/35057062`.

[3] Zahra Abdellah et al. "Finishing the euchromatic sequence of the human genome". In: *Nature* 431.7011 (2004), pp. 931–945. ISSN: 00280836. DOI: `10.1038/nature03001`.

[4] Deanna M. Church et al. "Modernizing reference genome assemblies". In: *PLoS Biology* 9.7 (2011). ISSN: 15449173. DOI: `10.1371/journal.pbio.1001091`.

[5] Evan E. Eichler, Royden A. Clark, and Xinwei She. *An assessment of the sequence gaps: Unfinished business in a finished human genome.* 2004. DOI: `10.1038/nrg1322`.

[6] Karen H. Miga et al. "Telomere-to-telomere assembly of a complete human X chromosome". In: *Nature* 585.7823 (2020), pp. 79–84. ISSN: 14764687. DOI: `10.1038/s41586-020-2547-7`.

[7] Sara Ballouz, Alexander Dobin, and Jesse A. Gillis. *Is it time to change the reference genome?* 2019. DOI: `10.1186/s13059-019-1774-4`.

[8] John Huddleston and Evan E. Eichler. *An incomplete understanding of human genetic variation.* 2016. DOI: `10.1534/genetics.115.180539`.

[9] Tayebeh Resaie. *Previewing GRCm39: assembly updates from the GRC.* NCBI. URL: `https://www.slideshare.net/GenomeRef`.

[10] Dongsheng Lu and Shuhua Xu. "Principal component analysis reveals the 1000 Genomes Project does not sufficiently cover the human genetic diversity in Asia". In: *Frontiers in Genetics* 4.JUL (2013). ISSN: 16648021. DOI: `10.3389/fgene.2013.00127`.

[11] Giorgio Sirugo, Scott M. Williams, and Sarah A. Tishkoff. *The Missing Diversity in Human Genetic Studies.* 2019. DOI: `10.1016/j.cell.2019.02.048`.

[12] Devon P. Ryan et al. "Mutations in a potassium channel (Kir2.6) causes susceptibility to thyrotoxic hypokalemic periodic paralysis". In: *Qatar Medical Journal* 9.2 (2000), pp. 70–72. ISSN: 02538253. DOI: `10.1016/j.cell.2009.12.024`.

[13] Thomas P. Moyer et al. "Warfarin sensitivity genotyping: A review of the literature and summary of patient experience". In: *Mayo Clinic Proceedings* 84.12 (2009), pp. 1079–1094. ISSN: 00256196. DOI: `10.4065/mcp.2009.0278`.

[14] Alice B. Popejoy and Stephanie M. Fullerton. *Genomics is failing on diversity.* 2016. DOI: `10.1038/538161a`.

[15] All of Us Research Program Investigators. "The "All of Us" research program". In: *New England Journal of Medicine* 381.7 (2019), pp. 668–676.

[16] Lucia A. Hindorff et al. *Prioritizing diversity in human genomics research.* 2018. DOI: `10.1038/nrg.2017.89`.

[17] Rachel M. Sherman et al. *Assembly of a pan-genome from deep sequencing of 910 humans of African descent.* 2019. DOI: `10.1038/s41588-018-0273-y`.

[18] Karen H.Y. Wong et al. "Towards a reference genome that captures global genetic diversity". In: *Nature Communications* 11.1 (2020). ISSN: 20411723. DOI: `10.1038/s41467-020-19311-w`.

[19] Ananyo Choudhury et al. "High-depth African genomes inform human migration and health". In: *Nature* 586.7831 (2020), pp. 741–748. ISSN: 14764687. DOI: `10.1038/s41586-020-2859-7`.

[20] Rachel M. Sherman and Steven L. Salzberg. *Pan-genomics in the human genome era.* 2020. DOI: `10.1038/s41576-020-0210-7`.

REFERENCES

[21] Genome Reference Consortia. *Human Genome Overview: Next assembly update.* Genome Reference Consortium. URL: https://www.ncbi.nlm.nih.gov/grc/human (visited on 07/14/2021).

[22] Karen H.Y. Wong, Michal Levy-Sakin, and Pui Yan Kwok. "De novo human genome assemblies reveal spectrum of alternative haplotypes in diverse populations". In: *Nature Communications* 9.1 (2018). ISSN: 20411723. DOI: 10.1038/s41467-018-05513-w.

[23] David Porubsky et al. "Fully phased human genome assembly without parental data using single-cell strand sequencing and long reads". In: *Nature Biotechnology* (2020). ISSN: 15461696. DOI: 10.1038/s41587-020-0719-5.

[24] Peng Zhou et al. "A pneumonia outbreak associated with a new coronavirus of probable bat origin". In: *Nature* 579.7798 (2020), pp. 270–273. ISSN: 14764687. DOI: 10.1038/s41586-020-2012-7.

[25] Wouter De Coster, Matthias H. Weissensteiner, and Fritz J. Sedlazeck. *Towards population-scale long-read sequencing.* 2021. DOI: 10.1038/s41576-021-00367-3.

[26] Amalio Telenti et al. "Deep sequencing of 10,000 human genomes". In: *Proceedings of the National Academy of Sciences of the United States of America* 113.42 (2016), pp. 11901–11906. ISSN: 10916490. DOI: 10.1073/pnas.1613365113.

[27] Elizabeth K. Ruzzo et al. "Inherited and De Novo Genetic Risk for Autism Impacts Shared Networks". In: *Cell* 178.4 (2019), pp. 850–866. ISSN: 10974172. DOI: 10.1016/j.cell.2019.07.015.

[28] Heng Li. "Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM". In: *arXiv preprint arXiv* (2013). ISSN: 2169-8287. DOI: arXiv:1303.3997[q-bio.GN]. URL: http://arxiv.org/abs/1303.3997.

[29] Guillaume Marçais and Carl Kingsford. "A fast, lock-free approach for efficient parallel counting of occurrences of k-mers". In: *Bioinformatics* 27.6 (2011), pp. 764–770. ISSN: 13674803. DOI: 10.1093/bioinformatics/btr011.

[30] Julie Hussin et al. "Age-dependent recombination rates in human pedigrees". In: *PLoS Genetics* 7.9 (2011). ISSN: 15537390. DOI: 10.1371/journal.pgen.1002251.

[31] Kelley Paskov et al. "Estimating sequencing error rates using families". In: *BioData Mining* 14.1 (2021). ISSN: 17560381. DOI: 10.1186/s13040-021-00259-6.

[32] G. David Forney. "The Viterbi Algorithm". In: *Proceedings of the IEEE* (1973). ISSN: 15582256. DOI: 10.1109/PROC.1973.9030.

[33] Peter Hall and Barbara La Scala. "Methodology and Algorithms of Empirical Likelihood". In: *International Statistical Review / Revue Internationale de Statistique* (1990). ISSN: 03067734. DOI: 10.2307/1403462.

[34] Johannes Alneberg et al. "Binning metagenomic contigs by coverage and composition". In: *Nature Methods* (2014). ISSN: 15487105. DOI: 10.1038/nmeth.3103.

[35] Adi Fledel-Alon et al. "Variation in human recombination rates and its genetic determinants". In: *PLoS ONE* 6.6 (2011). ISSN: 19326203. DOI: 10.1371/journal.pone.0020321.

[36] Brianna Chrisman et al. "Analysis of Sex and Recurrence Ratios in Simplex and Multiplex Autism Spectrum Disorder Implicates Sex-Specific Alleles as Inheritance Mechanism". In: *Proceedings - 2018 IEEE International Conference on Bioinformatics and Biomedicine, BIBM 2018.* 2019, pp. 1470–1477. ISBN: 9781538654880. DOI: 10.1109/BIBM.2018.8621554.

[37] Ambroise Wonkam. "Sequence three million genomes across Africa". In: *Nature* (2021). ISSN: 14764687. DOI: 10.1038/d41586-021-00313-7.